Anduril workflow framework

Kristian Ovaska, MSc
University of Helsinki
10.02.2012

Anduril workflow framework

- Workflow framework for scientific data analysis
- Scalability to complex analysis, heterogeneous data and large data sets
- Scripting language to construct workflows
- Reusing analysis code and (partial) workflows
- Open source
- Ovaska, et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Medicine, 2010, 2:65.

http://www.anduril.org

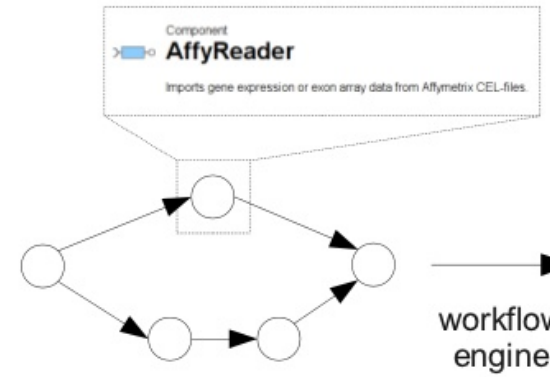
Design goals

- Motivation: bioinformatics high-throughput data analysis
- Also suitable for other "batch" analysis tasks
- Target users: researchers with some programming experience
- Apply software engineering to workflows: reuse, formal interfaces, testing

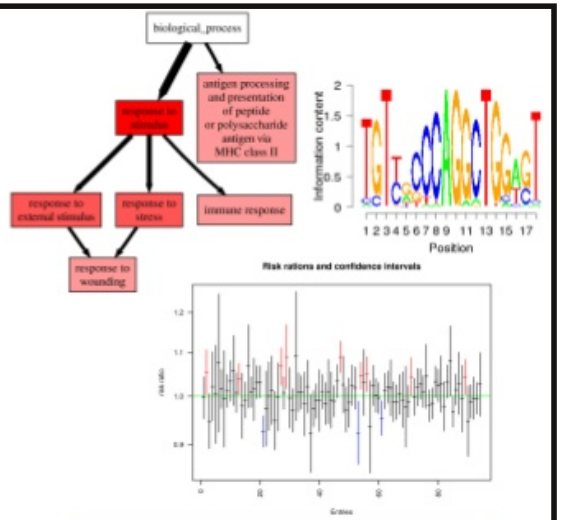
System overview

```
...
x1 = AffyReader(sampleNames)
if z > 5 { x2 = ComponentA() }
else     { x2 = ComponentB() }
x3 = CSVFilter(x1.expr)
...
```

AndurilScript interpreter



workflow engine



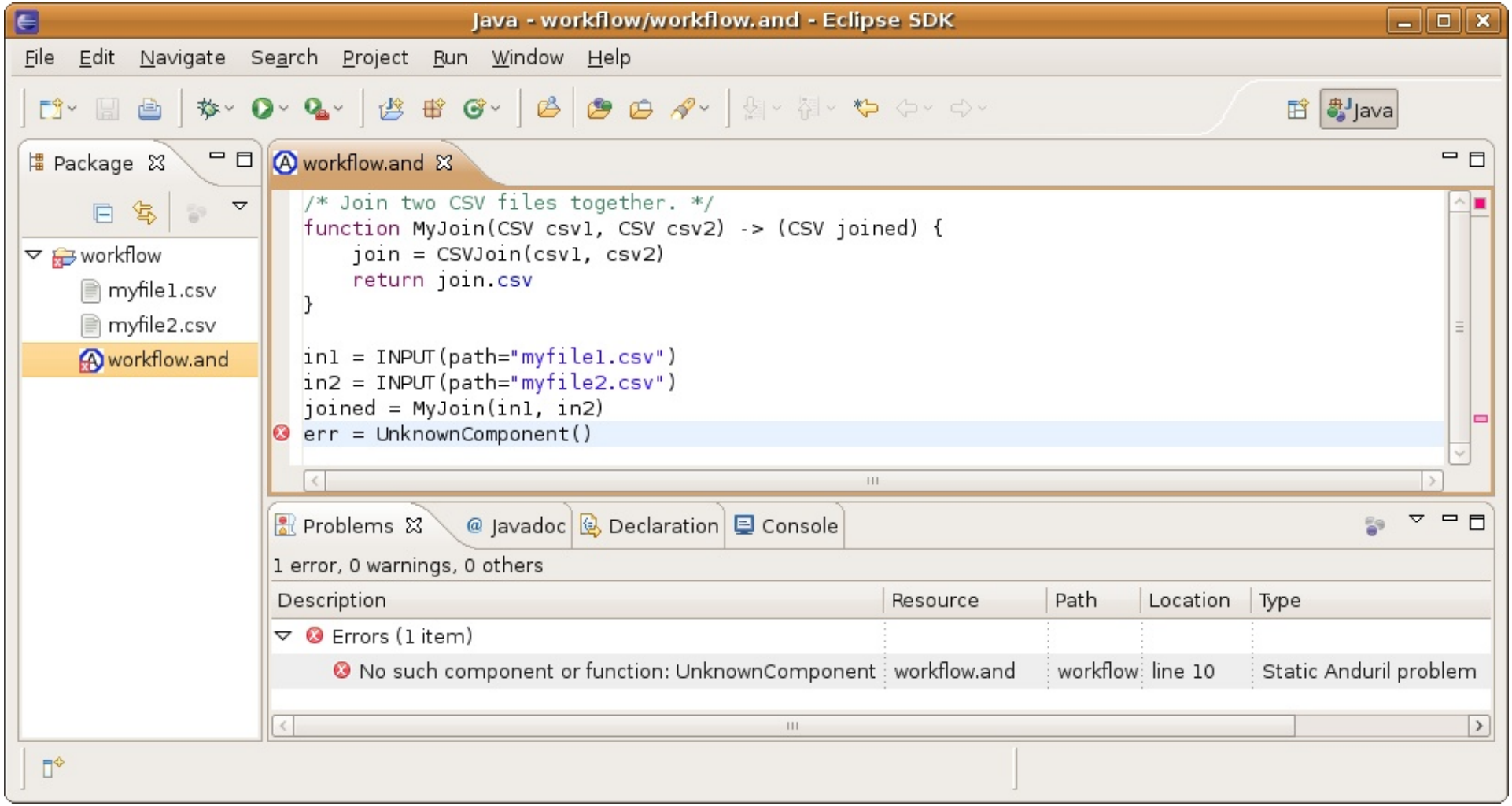
AndurilScript: Lightweight scripting language for workflow construction

Workflow: User-defined network of reusable components

Results and visualization: PDF, Excel, HTML

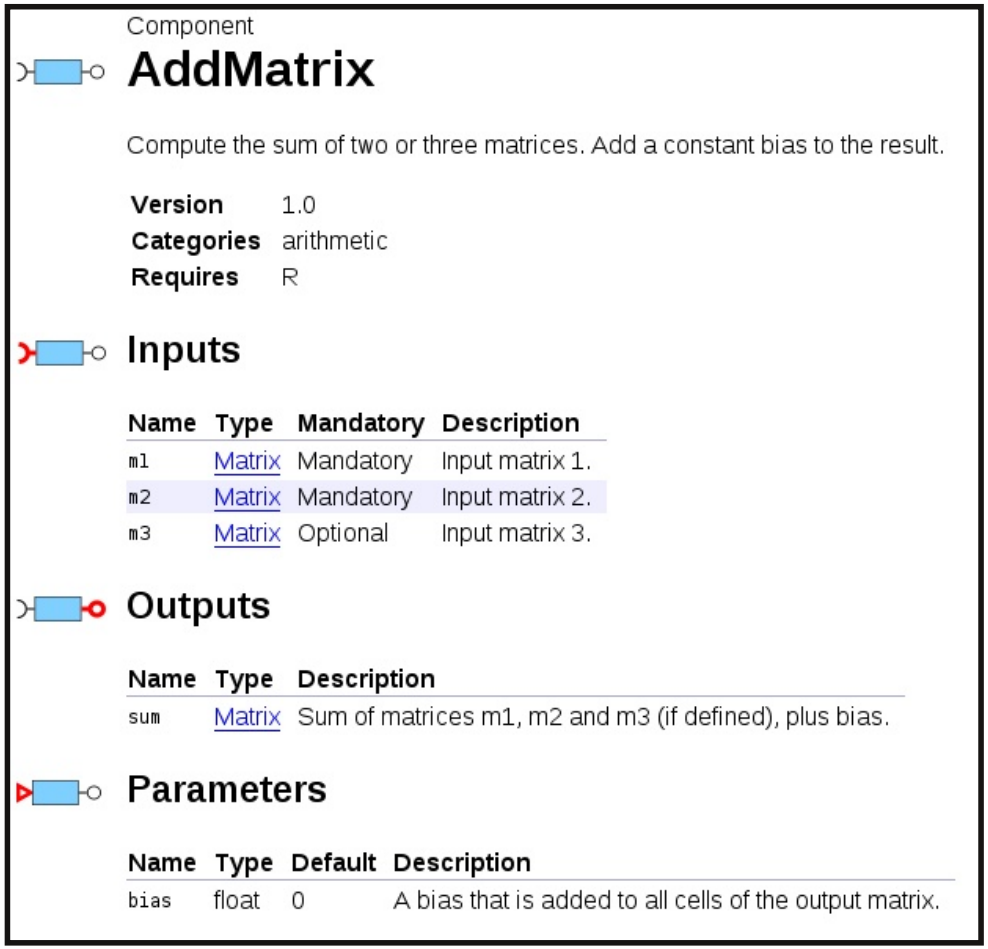
User interface

- Command line interface / Eclipse plugin



Component model

- Reusable code unit implementing a workflow step
- Executable program that reads and writes files
- Programming language independent: Java, R, Python, Bash, Perl, Matlab, Lua
- Rich type system for ports (subtyping, type parameters)



Component

AddMatrix

Compute the sum of two or three matrices. Add a constant bias to the result.

Version	1.0
Categories	arithmetic
Requires	R

Inputs

Name	Type	Mandatory	Description
m1	Matrix	Mandatory	Input matrix 1.
m2	Matrix	Mandatory	Input matrix 2.
m3	Matrix	Optional	Input matrix 3.

Outputs

Name	Type	Description
sum	Matrix	Sum of matrices m1, m2 and m3 (if defined), plus bias.

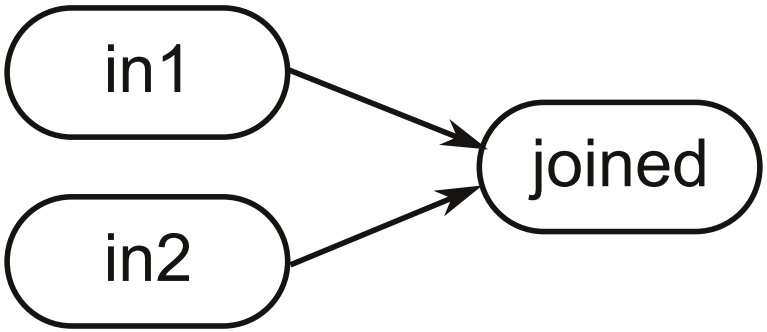
Parameters

Name	Type	Default	Description
bias	float	0	A bias that is added to all cells of the output matrix.

Workflow construction

- Workflow construction by programming: scalability

```
function Join(CSV x, CSV y) -> (CSV joined) {
    join = CSVJoin(x, y)
    return join.csv
}
in1 = INPUT(path="myfile1.csv")
in2 = INPUT(path="myfile2.csv")
joined = Join(in1, in2)
```



Workflow templates

- Flexibly reuse partial workflows
- Templates take parameters that enable conditional instantiation

```
function MyFunction(CSV x, optional Latex y,
          int n=1, boolean flag=true)
          -> (CSV out1, PDF out2) {
     if (n > 1 || flag || y == null) {
          // conditional task 1
     } else {
          // conditional task 2
     }
     return {"out1"=X, "out2"=Y}
}
```

Advanced features

- Looping in scripts: constructing large workflows
- Remote execution over SSH or other mechanism

```
table = ConstructTable()
for row: std.itercsv(table.csv) {
     subtask = SubTask(param=row.value,
          @host="remote", @name=row.id)
}
```

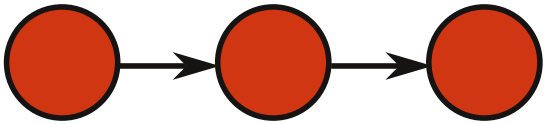
hosts.conf

```
HostID = remote
HostName = 192.168.1.1
IsSharedFileSystem = true
Slots = 4
```

Workflow engine

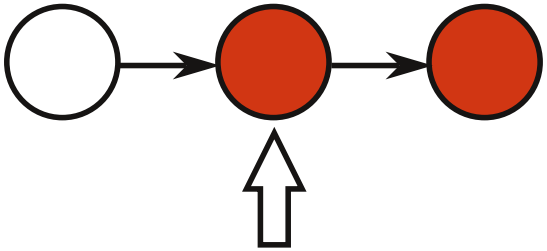
- Workflows have no loops (directed acyclic graph)
- Engine optimized for "agile analysis" with short cycles
- Execute heavy preprocessing steps only once

First run



Execute everything, cache results on disk

Second run



Execute only changed node and its dependants

Summary

- Anduril is a scalable workflow framework for users with some programming experience
- Visual final reports for non-computational people
- Optimized for iterative scientific data analysis
- Workflow construction by programming enables complexity scaling and reuse
- System used in >10 projects since 2010 publication

http://www.anduril.org

Acknowledgements

Computational Systems Biology Laboratory
Genome-Scale Biology Research Program
Institute of Biomedicine
Center of Excellence in Cancer Genetics
University of Helsinki
Dr. Sampsa Hautaniemi, PI

UNIVERSITY OF HELSINKI
FACULTY OF MEDICINE

Funding
Academy of Finland
Biocentrum Helsinki
Sigrid Jusélius Foundation
EU FP7
Helsinki Biomedical Graduate School